

A new method to reduce overestimation of thresholds with observational network data

George Berry,^{1*} and Christopher J. Cameron¹

¹Department of Sociology, Cornell University,
368 Uris Hall, Ithaca NY 14853, USA

*To whom correspondence should be addressed; E-mail: geb97@cornell.edu.

Abstract

Networks of interdependent nodes support phenomena such as epidemics, product adoption, cascading failure, ecosystem collapse, congestion, and bandwagon effects. We consider the problem of using observational data to estimate the sensitivity of individual nodes to the activation of their network neighbors. We prove that—in the case of binary activation decisions—activation thresholds are impossible to correctly measure for some nodes in virtually all contagion processes on complex networks. This result holds even when each step of the process is observed. Measurement error always produces an overestimate of a node’s true activation threshold. We develop a condition for determining which node thresholds are correctly measured and demonstrate that modeling activation thresholds as a function of node-level factors reduces the error compared to existing approaches.

1 Introduction

A major contribution of contemporary social science is the study of social contagion, or social phenomena which spread through social networks [14, 23, 24, 6]. Although an independent branch of sociological research has developed to categorize these phenomena [20, 25, 26, 23, 10, 27], such analyses

are technically similar to studies of epidemics [19, 28, 5], power grid failures [7], bank panics [12], ecosystem collapse [2], online virality [3], adolescent peer effects [13], and cascading failure [4, 16].

A useful abstraction for describing contagion on a complex network is the *threshold model* [14, 11, 18, 6]. In this model, each node has an *activation threshold* corresponding to the number of neighbors that must activate before the focal node activates. These models have yielded wide theoretical insights across the social and information sciences, but scholars have struggled to apply them empirically¹.

In this article, we describe a fundamental property of the threshold model of contagion: in most cases, recording the number of active neighbors (*exposure*) when a node activates (the *exposure-at-activation-time* rule) overestimates many node thresholds. We specify the conditions for which this bias is certain. We generalize these results from threshold models of contagion to empirical models of contagion that do not explicitly employ the concept of a threshold.

The problem of threshold measurement is potentially very serious yet under-appreciated. In contrast to the well-known reflection problem for studying peer effects [21], the *measurement problem* we identify cannot be fixed by alternative model specification or by collecting more fine-grained data. To our knowledge, we are the first to identify this problem and develop a solution. We describe a *correct measurement condition* which identifies which nodes in an empirical process are correctly measured. Importantly, researchers must observe node activation status before and after activation time for correct measurement to be possible.

We propose a method that uses node-level information and the correct measurement condition in order to dramatically reduce the measurement error of using the exposure-at-activation-time rule. This approach allows modeling how node characteristics relate to activation thresholds, which facilitates addressing the cascade prediction problem [8]. Our analysis relies on a general framework with implications for fields such as epidemiology, physics, ecology, sociology, political science, economics, and computer science.

¹See [25, 26] for notable exceptions. In addition, a footnote in [25] p. 75 recognizes a subset of the issues we raise here—namely, that a lag in adoption can allow time for additional network neighbors to adopt.

2 Intuition

Consider a simple laboratory experiment where we wish to learn an individual i 's activation threshold for purchasing a new product. We could conduct this experiment by showing i the product, indicating that no peers have activated, and ask whether i would purchase the item. Then, we show i the product with a single friend adopting and ask again, repeating this process until i answers in the affirmative. Assume for the sake of example that there is no cumulative effect of repeat exposures, and that the decision to purchase is determined solely on the number of peers who have purchased.

Assume i activates with three active neighbors. Since we know i was inactive with two active neighbors and active with three active neighbors, we correctly measure i 's activation threshold as three. This suggests a *correct measurement condition*: to correctly measure node i 's activation threshold h_i , the node must update and remain inactive with exposure $h_i - 1$ and then update and activate with exposure h_i . If this condition is met, exposure at activation time is the node's threshold.

With observational data, we cannot reliably replicate this experiment. We cannot control when i updates relative to its neighbors, nor can we control the connections among i and its neighbors. Perhaps i updates first with zero active neighbors, and then again with 10 active neighbors. If i activates at exposure 10, we only know that i 's threshold lies in the interval $(0, 10]$. This is the core intuition of our main result: empirical observation of contagions on networks reveals threshold intervals, not point estimates. When intervals have size exactly equal to one, i 's exposure at activation time is its threshold. Otherwise, uncertainty exists about where i 's threshold lies in an interval.

The exposure-at-activation-time rule used in past research takes the maximum of this interval, leading to a systematic upward bias in threshold estimation. To understand the potential severity of this problem, consider a fully connected graph of n nodes where one node has a threshold of zero and the remaining nodes each have threshold of one. If the nodes update one by one in random order, the observed exposures at activation time will be given by the sequence $0, 1, 2, \dots, n - 1$. Only the first two nodes to activate will have an exposure at activation time that reflects their true thresholds. The remaining $n - 2$ nodes will activate with exposure higher than their true thresholds, with a maximum measurement error of $n - 2$. In general, if a node i has a non-negative threshold, the upper bound on measurement error for i is the degree of i .

3 Proof of threshold immeasurability

Assume that a researcher observes a contagion process unfold on a network and does not have control over the graph structure, threshold assignment, node update ordering, and node activation delays. Also assume that when a node activates, its activation is publicly known to its neighbors and researchers. This is the default setting for studying social contagion [1, 10] and the case we consider here (additional cases are provided in the appendix.).

A process is *threshold measurable* if we are guaranteed to correctly measure all node thresholds according to the correct measurement condition, regardless of the values of the elements that are out of the researcher’s control. In other words, all threshold intervals must be equal to one with certainty. To prove that threshold measurability does not hold for a contagion process, we provide a counter-example using the elements a researcher does not control. If one or more nodes is incorrectly measured, then the counter-example is valid and the contagion process is not threshold measurable.

Fig. 1-B provides a valid counter-example, proving that contagion processes are not threshold measurable when the researcher does not control the graph structure, threshold assignment, node update ordering, and node activation delays. In this counter-example, we observe an interval $(0, 2]$ for a node’s threshold, which violates the correct measurement condition. The exposure-at-activation-time rule would assign this node a threshold of two when its true threshold is one.

Contagion on a large graph is composed of contagion on many small graphs. This fact implies that we should not expect to correctly measure all node thresholds in similar processes on larger graphs. The rate and size of measurement error are empirical questions beyond the scope of this proof (below we assess these with simulation).

4 Implications for estimates of contagion

The upward bias in threshold measurement has implications for models of social contagion. We focus on two such models: 1) the probability of first activation at a given exposure level; 2) a regression model estimating the cumulative probability of being active at a given exposure level. By “first activation” we mean the lowest exposure value at which a node activates, while “cumulative” denotes the probability of being active at a given exposure

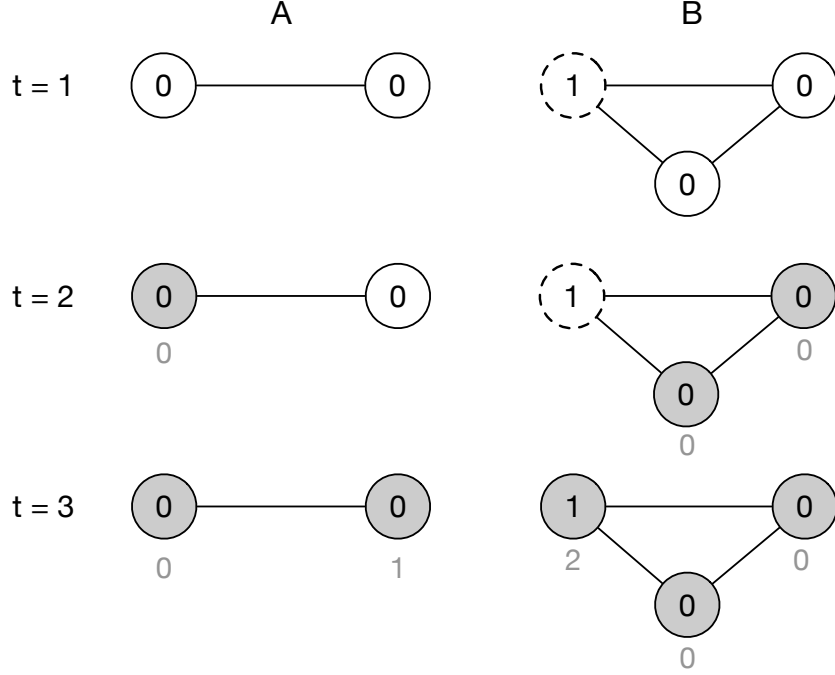


Figure 1: Graphs over three time periods. Transparent nodes are inactive, gray nodes are active. A dashed border indicates that a node has updated but remains inactive. The number on the node indicates the true activation threshold, while the number below the node indicates the measured threshold. A node is correctly measured when these two numbers are equal. We assume no activation delays. In A, the right node has a measured threshold of one when its true threshold is zero. In B, the leftmost node is measured with a threshold of two when its true threshold is one. In A, updating both nodes at once (synchronous updating) allows correct measurement of both nodes. Adding a third node, as in B, still results in measurement error. In figure B, both synchronous and asynchronous updating lead to measurement error. In fact, any strategy of updating nodes in B produces measurement error.

level.

Note that a contagion process yields measurements of the form (a_i, r_i) , or the activation state ($a_i \in \{0, 1\}$) and exposure count (r_i , a non-negative integer) for node i . We will denote a general exposure level as r in the case where we wish to discuss the probability of first activation at a given exposure level, for instance $r = 3$.

4.1 Probability of first activation

A common estimate of contagion is the probability of first adoption at exposure r [22]. This is given by $|r_{\text{first-active}}|/(|r_{\text{first-active}}| + |r_{\text{inactive}}|)$, where $|r_{\text{first-active}}|$ indicates the cardinality of the set of nodes first activating at r , and $|r_{\text{inactive}}|$ indicates the cardinality of the set of nodes still inactive at r . Measurement error effects both the numerator and denominator through $|r_{\text{first-active}}|$. We can easily see that, holding $|r_{\text{inactive}}|$ constant, reducing $|r_{\text{first-active}}|$ (the count of first activators) will reduce the value of the expression, and increasing $|r_{\text{first-active}}|$ will increase the value of the expression.

Consider the case of a true threshold of three being over-estimated with a value of 10. We will reduce the estimated probability of first activation at exposure three, and increase the estimated probability of first activation at exposure 10. Because measurement error always inflates thresholds, this pattern of under-estimating activation probabilities at low exposure and over-estimating at high exposure can be considered quite general, although specifics will vary by case. Table 1 describes the severity of this problem for a representative simulation run.

4.2 Regression estimate of cumulative activation probability

Consider the regression case where we wish to estimate the cumulative probability of activation at a given exposure level, written $a_i = \beta r_i + \epsilon_i$, where a_i is activation status, r_i is i 's exposure, β relates r_i to the outcome, and ϵ_i is noise uncorrelated with r_i . Logit and probit are more common in practice, but we discuss a linear probability model for simplicity.

Consider the effect of measurement error on this model. Some nodes will have their first activation moved from lower to higher exposure values, e.g. from $r = 3$ to $r = 10$. This will result in the estimated β being

smaller than its true value, since nodes appear to first activate at higher exposure levels. We express measurement error as $r_i = h_i + m_i$, where r_i is the measured threshold, h_i is the true threshold, and m_i is measurement error. $E[m_i] > 0$ and $m_i \geq 0$ for all i . In addition, m_i may possibly be correlated with h_i . These facts render inapplicable errors-in-variables models which assume mean-zero error uncorrelated with the independent variable in question [9].

5 Measurement strategy to reduce bias

The proposed measurement strategy is based on the fact that empirical observation of contagions on networks reveals intervals rather than point estimates. We propose leveraging the width of the threshold intervals to identify correctly measured nodes. Using observational data, each activated node can be associated with two values: 1) the highest exposure prior to activation, and 2) the exposure at activation. These two values establish upper and lower bounds for an interval that contains the node’s true threshold. The subset of nodes with their threshold bounded by an interval of width one constitute the set of correctly measured nodes. The proportion of correctly measured nodes and the interval widths for incorrectly measured nodes in turn indicates the extent of the measurement problem for the specific network and contagion process.

If node attributes are known and correlated with threshold values, the subset of correctly measured nodes can be used to model the relationship between attributes and threshold, using the correctly measured set. The model can then be used to predict the thresholds for nodes outside of the correctly measured set.

5.1 Evaluation through simulation

We used simulations designed to assess the severity of the measurement error and the efficacy of our proposed error-reduction technique. We focus on a power law with clustering graph [15] with clustering parameter 0.1, mean degree 12, 1000 nodes, and 1000 replications. Additional graph types are discussed in the appendix². This type of simulated graph exhibits high het-

²Results were similar across all simulations we conducted.

erogeneity of node degree while building in clustering that is typical of social and information networks.

In each replication, we drew node thresholds h_i from the function $h_i = 5 + 3x_i + \epsilon_i$, where x_i is a node-level covariate drawn from a $\mathcal{N}(0, 1)$ and ϵ_i represents a node-level idiosyncratic error also drawn from $\mathcal{N}(0, 1)$. This threshold generation procedure yields a normally-distributed *threshold distribution*. We recorded the largest exposure before activation and the exposure at activation time, allowing us to apply the correct measurement condition defined above.

On average 720 nodes activated and 113 were correctly measured. The observed distribution of activated nodes using the exposure-at-activation-time rule differs from the true distribution of thresholds (Fig. 2). The observed distribution is shifted to the right and has a long tail that is not present in the true distribution. The long tail results from the presence of high degree nodes that update after their neighbors due to chance, inflating their measured thresholds.

In the correctly measured set (Fig. 3), correctly measured nodes tend to have values below the mean of the threshold distribution. This suggests that there is selection on the idiosyncratic error, since a node with a more negative error term is more likely to have a low threshold, and therefore more likely to be measured³. This implies that using the correctly measured subset to model h_i as a function of x_i will not work perfectly because $E[\epsilon_i] < 0$ in the correctly measured subset.

We assess the usefulness of using information contained in x_i to model node thresholds with three different models. First, we use the correctly measured subset (on average 113 out of 1000 nodes) to estimate an ordinary least squares (OLS) model of form $h_i = \alpha + \beta x_i + \epsilon_i$, where h_i is i 's true threshold. Note that due to selection on the error, we do not expect this model to consistently estimate α and β . Second, we use the entire set of activated nodes (on average 720 out of 1000 nodes) to estimate an OLS model of form $r_i = \alpha + \beta x_i + \epsilon_i$, where $r_i = h_i + m_i$ is the true threshold plus measurement error. Rewriting this equation yields $h_i = \alpha + \beta x_i + (\epsilon_i - m_i)$. We do not have a reason to expect $E[\epsilon_i - m_i] = 0$ among the activated subset, and we know that $E[m_i] > 0$, suggesting that the error term among

³We attempted to model this selection using structural factors, such as node degree, closeness, betweenness, and eigenvector centrality. These did not help in simulations we conducted.

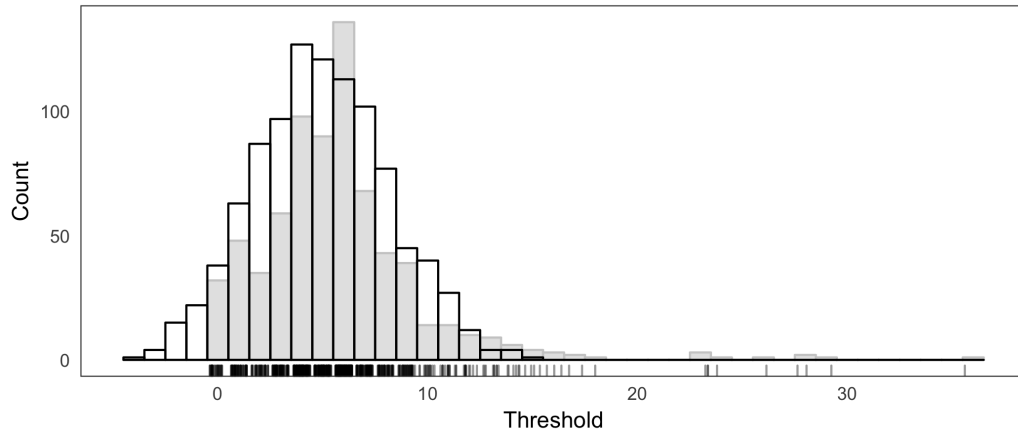


Figure 2: Results from a representative simulation with 1000 nodes, of which 720 activated in this run. Transparent bars are the true threshold distribution for all nodes, whereas gray bars are thresholds as measured by the exposure-at-activation-time rule. The maximum measured threshold is 36 (node true threshold = 15). The largest relative error is an 800% over-measurement, due to a node with true threshold 2 being measured with threshold 16. This upward measurement error results in the long tail we see here.

the activated subset is also negative on average. Third, we use the exposure-at-activation-time rule for the activated nodes (on average 720 out of 1000 nodes) and take $RMSE(h_i, r_i)$.

Fig. 4 presents the RMSE for these three modeling approaches. Using only the correctly measured subset (11.3% of the data on average) consistently produces the lowest RMSE results. In addition, we obtain a prediction for all nodes in the graph, which is not possible using the exposure-at-activation-time rule. Based on this result, we suggest using the correctly measured subset to model node thresholds with individual-level covariates.

This model-based approach to estimating thresholds has the additional advantage of providing a new tool to address the cascade prediction problem [8]. The first k correctly measured thresholds can be used to predict thresholds for all nodes in the graph. Fig. 5 presents the RMSE predicting thresholds using the first k correctly measured nodes. Using the first 40 correctly measured nodes improves upon the exposure-at-activation-time rule, while using the first 70 provides a good estimate. This application of our modeling strategy facilitates predicting node thresholds from partial contagion data, which in turn allows a prediction of cascade depth.

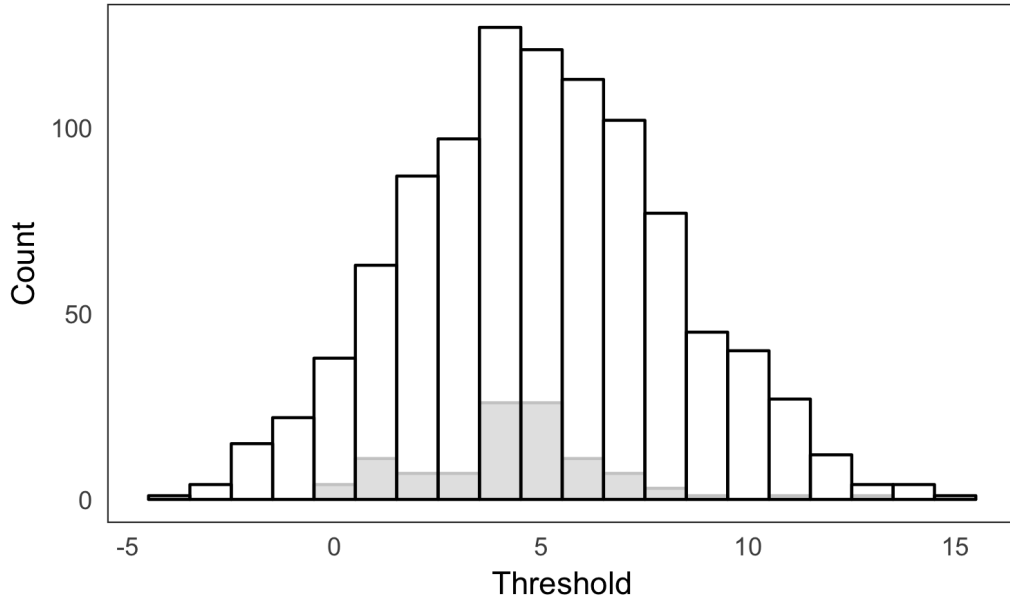


Figure 3: Results from a representative simulation with 1000 nodes, of which 105 were correctly measured. Transparent bars are the true threshold distribution of all nodes, whereas gray bars are the correctly measured thresholds. Note that we tend to correctly measure nodes at the lower end of the true threshold distribution, indicating selection on the error. In other words, a node with a negative error term will be more likely to be correctly measured.

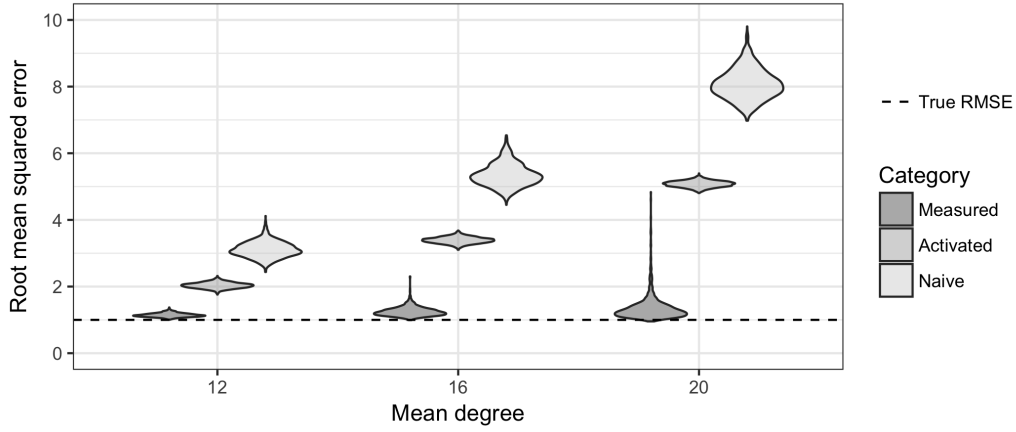


Figure 4: Across 1000 runs of the simulation, we evaluate the RMSE of three different methods for recovering node activation thresholds: 1) *Measured*: estimate OLS model with the correctly measured subset and predicting for all nodes; 2) *Activated*: estimate OLS model with all activated nodes and predict for all nodes; 3) *Naive*: use the exposure-at-activation-time rule. The correctly measured subset consistently performs the best despite using by far the least amount of data. On average, 720 nodes activate and 113 are correctly measured. In other words, estimating a model with 10% of the nodes in the graph outperforms two models that use 70% of the nodes in the graph.

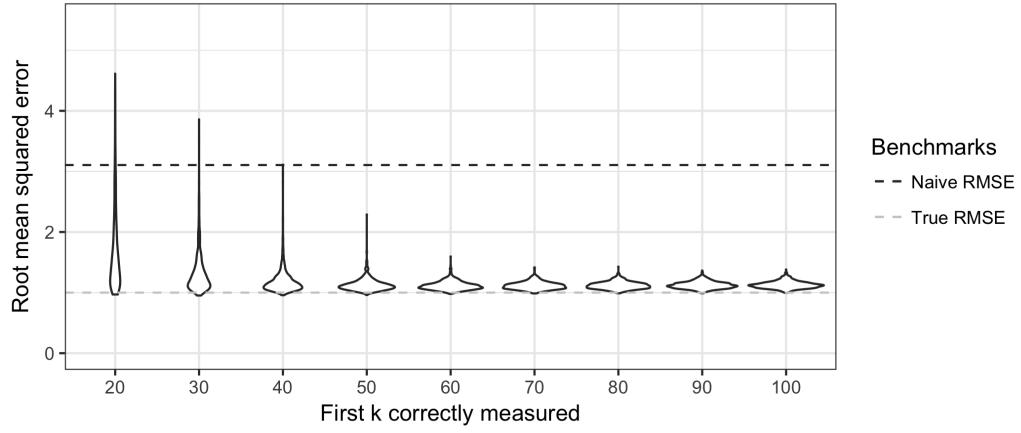


Figure 5: Across 1000 runs of the simulation, we use the first k correctly measured nodes to estimate a model and then predict thresholds for all nodes. We plot the RMSE using violin plots, and compare to the true RMSE (gray line) and the “naive RMSE” using the exposure-at-activation-time rule (black line). We see that by $k = 40$, we always do better than the naive rule, and by $k = 70$, we obtain low-RMSE estimates across all runs.

5.2 Practical guidance

Several practical difficulties arise when attempting to assess the severity of measurement error and apply the modeling strategy proposed here. Most importantly, the data available must have records of node update behavior apart from activation. Appropriate indicators of updating or experiencing exposure would be login times or engagement metrics like click-through or viewing. This requires multiple observations for individual nodes over time. Data sets that do not contain information on node update behavior prior to activation are not suitable for the modeling procedure that we have suggested.

Practically speaking, a cross-sectional snapshot of a contagion process does not lend itself to the modeling strategy we propose. Snapshots yield only a mixture of inactive nodes and upper bounds of threshold intervals, but will not measure any single node’s threshold interval. We leave work adapting our techniques to cross-sectional contagion snapshots to future research.

Determining when a node updates presents a second practical difficulty that will depend on the contours of the research site. As an illustration, consider the case of estimating the number of neighbors that must adopt a hashtag on Twitter before an individual adopts. We can make the assumption that a node “updates” when sharing a public status, which yields multiple observations for each node across time. This allows the construction threshold intervals for nodes, and application of the correct measurement condition.

Finally, there is a subtle theoretical difficulty concerning the model of node thresholds presented in the previous section. If node thresholds can be thought of as “negative” (as we model above), all nodes which activate with zero active neighbors should be excluded from the correctly measured set. In this case, there is no update data before a node activates to establish a threshold interval. We suggest excluding zero cases from the correctly measured set, since negative thresholds may represent nodes which are particularly inclined to adopt the contagion.

6 Conclusion

As data collection methods improve across the sciences, we have access to increasingly fine-grained data about contagion processes. These observational data now make it possible to study contagions by estimating thresholds. We

show that unprecedented opportunities for empirical assessment pose previously unrecognized measurement problems, even when models are correctly specified and time-stamped activity data are available for every node in the network. The error becomes even more pronounced when nodes are surveyed in waves or when networks are dense and thresholds are low.

Fortunately, the problem can be addressed. By leveraging the correct measurement condition we have identified, researchers can determine the scope of the measurement problem for a particular case. By using the suggested modeling approach, researchers can build a model for node thresholds which reduces measurement error.

Models of node thresholds can be used to study complex contagion [6] empirically using observational data. If a model of thresholds shows that most nodes need social reinforcement before activation (if most thresholds are greater than one), then this provides an argument for a contagion being complex. This has ramifications for how we expect contagion to spread and how to approach practical problems such as the influence maximization problem [17].

We also suggest that this model-based methodology can be used to identify individuals who are particularly susceptible to contagion. For instance, in online settings, users participate in many contagion processes at once (e.g. they adopt many hashtags). By estimating a model for each hashtag, we can obtain user-level predictions for adopting many different types of behaviors, allowing inference about characteristics of individuals.

Finally, future work is needed to refine the modeling approach we have presented here. While we consider only correctly measured thresholds for our model, there are many thresholds which fall within a narrow interval. Future work should address how to better incorporate the information contained in small intervals to predict thresholds for all nodes.

7 Appendix

7.1 Simulation details

We simulated contagion processes on graphs generated by a power-law with clustering (clustering parameter = 0.1) and a Watts-Strogatz small-world network (rewriting probability = 0.1). We used mean degrees of 12, 16, and 20. We used error standard deviations in the threshold distribution of

0.5, 0.8, 1.0, 1.5, 2.0. All graphs had 1000 nodes, and all parameter sets were replicated 1000 times. Results were qualitatively similar to those in the main text for all simulation parameterizations with caveats discussed below.

To simulate contagion dynamics, we ran a series of iterations over the set of unactivated nodes (randomly permuted). We terminated a simulation run when all nodes were active, or when we updated each inactive node and observed no activations occur. This simulates a continuous time process where nodes update one at a time with no activation delay and perfect information about neighbor activation status.

We also attempted to use a simultaneous updating procedure where we updated all nodes each wave, but found that it offered extremely poor conditions for measuring thresholds. Certain parameterizations also offered poor conditions for threshold measurability. Such parameter sets were characterized by many low thresholds, leading to “rapid” activation. For instance, high degree and high error standard deviation do not facilitate measuring thresholds, since nodes have many neighbors with very low thresholds, driving down the probability of any instance replicating a laboratory experiment.

All code used to replicate our analysis is available in a GitHub repository. Researchers should use Python 3 with NetworkX, Pandas, Scikit Learn, Numpy. Graphs are generated with R using ggplot2 and dplyr.

7.2 Formalization

We lay out a full proof here. This follows similar logic to the proof in the main text, but covers additional cases and formalizes the notion of a “contagion process”. We begin by noting that the parameter space of contagion processes is very large. In particular, the following items are allowed to vary and we wish to make as few assumptions as possible about them:

1. The graph structure
2. Node thresholds
3. Node update ordering
4. Node activation delays

In addition to these features of contagion processes, there is a second set of factors that are important for the study of contagion: researcher knowledge.

For example, perhaps nodes have activation delays, so that threshold satisfaction and public activation are decoupled. This raises the important issue of which information the researcher has access to. In the case of internet routers, we may know an activation delay explicitly; in the case of humans adopting linguistic conventions, it is unlikely that we know such activation delays even if they exist.

From this discussion, we find that the following two considerations are also important for measuring contagion processes:

1. Whether an activation delay exists in the first place
2. If an activation delay does exist, whether the researcher has information about a node's private threshold satisfaction status before public activation

We will keep these factors in mind.

7.2.1 Notation

We lay out our full notation here for clarity. Individual nodes in a graph are indexed with i , and time periods are indexed with t . We use subscripts for individuals i and superscripts for times t .

- $G = (V, E)$ —the graph structure composed of sets of nodes V and edges E . Vertices are labeled $\{1, 2, 3, \dots, |V|\}$.
- H —the vector of length $|V|$ containing thresholds for each vertex, with thresholds for node i denoted h_i .
- U —the node update ordering. u^t , specifies the set of nodes that update simultaneously at time t . u_i denotes the set of all time periods where node i updates.
- Δ —the vector of length $|V|$ containing node activation delays, with δ_i corresponding to the activation delay for node i .
- a^t —the vector of length $|V|$ indicating which nodes have publicly activated at time t . If i has publicly activated at t , then we write $a_i^t = 1$, else $a_i^t = 0$.

- s^t —the vector of length $|V|$ indicating which node thresholds have been met at time t , but that have not necessarily publicly activated due to an activation delay. We write $s_i^t = 1$ if i 's threshold has been satisfied at t , otherwise $s_i^t = 0$.
- r^t —the vector of length $|V|$ indicating each node's exposure at time t . In other words, how many active neighbors does i have at time t .
- ω^t —the *state of the world* at time t . In other words, a tuple $\omega^t = (G, H, U, \Delta, e^t, s^t, a^t)$. The elements without a time subscript are fixed by assumption (e.g. each node has a fixed activation delay specified by $\delta_i \in \Delta$). The update ordering U is fixed for the contagion process, and at time t nodes u^t update. We write e^t, s^t, a^t to indicate that these elements are “dynamic”. At time t , we check the nodes specified by u^t and update e^t, s^t , and a^t accordingly.
- Ω —*states of the world*. This is shorthand for indicating all possible ω^t .

7.2.2 Dynamics

We conceive of contagion dynamics very simply. At each time t , we check the nodes to be updated at this time period $i \in u^t$. We record exposure r_i^t (the count of publicly activated neighbors of i) and, if $r_i^t \geq h_i$ set $s_i^t = 1$. This indicates that i 's threshold has been satisfied, but i is not yet publicly active. We then schedule public activation for time $t + \delta_i$, meaning i will publicly activate initially at $t + \delta_i$. $a_i^{t+\delta_i} = 1$ while $a_i^t = 0$. If $\delta_i = 0$, then $a_i^t = 1$.

We update all $i \in u^t$ simultaneously, meaning that the exposure counts r_i^t are those from before the beginning of the update step, rather than dynamically adjusting as we update each consecutive $i \in u^t$.

This update method corresponds to an assumption that each neighbor has identical influence on i 's activation, which can be relaxed without changing our substantive results here. Continuous-time processes are represented as those where exactly one node updates per update step.

7.2.3 Assumptions

These assumptions are mostly technical. The important one is what we call *granularity*: that the researcher sees every state of the world ω^t that is

realized by the contagion process. This means that we can't fix the issues we discover by sampling the process in a more fine-grained way.

Assumption 1. *Non-intervention:* G and H are exogenous. In other words, the researcher does not have control over them.

Assumption 2. *Ignorance:* The researcher does not in advance know any $h_i \in H$.

Assumption 3. *Granularity:* The researcher knows every ω^t that the contagion process achieves, subject to observational constraints such as not knowing H .

We also make two technical assumptions that make the proof below more straightforward:

Assumption 4. *Saturation:* All nodes in V activate eventually

Assumption 5. *Consistency:* G, U, H , and Δ are consistent with Assumption 4.

Practically speaking, the contagion process does not have to saturate for our result to apply. Simply take the subgraph of eventually activated nodes, then Assumption 4 and Assumption 5 apply and our results hold.

7.2.4 Definitions

With this notation and these assumptions, we can formulate our key concept: the *correct measurement* condition for a single node. We then call a contagion process *threshold measurable* if we are guaranteed to measure all node thresholds correctly.

First, we have two definitions for completeness.

Definition 1. An element is **exogenous** when the researcher does not have control over it. G and H are always exogenous, while U and Δ may or may not be exogenous.

For instance, we say that G is exogenous because the researcher does not determine the graph structure. U and Δ may or may not be exogenous, however. If an element is exogenous, we can think of nature as having control over the element to make threshold measurement as difficult as possible.

Definition 2. A *contagion process* is a combination of the mechanics of contagion in ω^t , the exogeneity status of U and Δ (G and H are exogenous by assumption), and the researcher’s ability to observe Δ and s^t .

The mechanics of a contagion process are encoded in ω^t as defined above. However, a contagion process also contains information about which elements are exogenous, and which elements the researcher can observe. Elements U and Δ may be exogenous in some processes, and in others the researcher may determine them. Likewise, the researcher may observe Δ or s^t in some processes, but not in others.

We now turn to defining our core concepts of correct measurement and process-level threshold measurability. For the first, there are two cases which are slightly different, depending on whether the researcher can observe s^t or only a^t . Essentially, s^t is “better” than a^t for measuring thresholds, so we should use s^t if available, otherwise we default to a^t . For the sake of simplicity, assume for the moment we can only observe a^t .

Definition 3. Node i ’s threshold is **correctly measured** if, for some $t, t' \in u_i, t' > t$, we have $r_i^{t'} - r_i^t = 1$ and $a_i^{t'} = 1$ while $a_i^t = 0$. In this case, the node’s threshold $h_i = r_i^{t'}$.

Note: If we can observe s^t , simply replace references to a^t with s^t throughout.

In words, Definition 3 says that a node’s threshold is correctly measured if we observed it unactivated at some time t , and activated at some later time t' , while its exposure at time t and t' differs by exactly 1. In this case, the contagion process simulates an experiment and allows us to correctly measure h_i , which is equal to the exposure at activation time $r_i^{t'}$.

This node-level condition allows formulating a straightforward process-level condition.

Definition 4. A contagion process is **threshold measurable** if we are guaranteed to correctly measure all thresholds, regardless of the values of the exogenous elements.

In other words, if (G, H, U, Δ) are exogenous, then we must show that regardless of the values of these elements, all nodes are guaranteed to be correctly measured. To prove it is not threshold measurable, we need only provide a single counterexample. This setup allows us to avoid making assumptions about graph generation processes, threshold distributions, and so forth.

While this may seem strict, our argument is simply that with observational data, (G, H, U, Δ) are actually exogenous. The space of (G, H, U, Δ) is vast and single elements are the subject of entire bodies of research. We leave it to future research to determine which assumptions about this space lead to better or worse prospects for measuring thresholds.

7.2.5 Proof strategy

Our proof strategy should be clear by now: state the features of the contagion process, assume it is threshold measurable, and then attempt to find a counterexample which contradicts this assumption. In one case, we will see that we provide a constructive proof that such a process is measurable.

7.2.6 Default case

Theorem 1. *If (G, U, H, Δ) are exogenous, the process is not threshold measurable, even if we can observe Δ and s^t .*

Proof. Since we play the role of “nature” setting the exogenous parameters, we have control over (G, U, H, Δ) . Assume that $V = (i, j, k)$, $E = ((i, j), (j, k), (k, i))$, and $h_i = h_j = 0$, $h_k = 1$. Let $U = (\{k\}, \{i, j\}, \{k\})$ so that k updates at time 1, i and j update at time 2, and k updates again at time 3. Let $\Delta = (0, 0, 0)$, indicating no lags in public activation.

Then the contagion process plays out as follows: at time 1, k checks and sees that it has exposure 0, which is less than its threshold of 1, and so $a_k = 0$. At time 2, i and j activate since they have threshold 0 and no activation lag. Then at time 3, k checks and sees that $r_k = 2$ while $h_k = 1$, and so activates immediately. Since k ’s two updates produce $r_k^1 = 0$ and $r_k^3 = 2$ while $a_k^1 = s_k^1 = 0$ and $a_k^3 = s_k^3 = 1$, Definition 4 is not satisfied and the process is not threshold measurable. In particular, all we know is that k ’s threshold lies in the interval $(0, 2]$

□

Theorem 1 is our main result, and can be seen graphically in Fig. 6. When (G, U, H, Δ) are exogenous, it is trivial to create small graphs on which nodes are not correctly measured. This implies the process is not measurable. Since large contagion processes are in a sense composed of small contagion

processes, our ability to find small structures that violate threshold measurability implies that we should not expect observational contagion processes to be threshold measurable.

Note that this result holds even if we know activation delays and private threshold satisfaction statuses. This happens primarily because we cannot know a node's threshold satisfaction status if the node itself has not updated, and we don't control node updates in this process.

7.2.7 Case where researcher controls node update ordering, but does not know private threshold satisfaction status or activation delays

We can immediately strengthen this result by asking the question: how sensitive is Theorem 1 to the specific update ordering U ? G or H may be exogenous, but perhaps researchers can allow nodes to update selectively. We show that if the researcher does not observe s^t and Δ , even having control over update ordering U does not allow threshold measurability.

Theorem 2. *Assume (G, H, Δ) are exogenous, and that the researcher does not observe s^t and Δ . Then the contagion process is not threshold measurable.*

Proof. We play the part of nature and provide a (G, H, Δ) that, for any U provided by the researcher, fails to be threshold measurable when we don't observe s^t and Δ . The intuition here is that even if the researcher controls U , not knowing the exogenous Δ and only observing public activations a^t makes threshold measurement impossible.

We formalize this by defining an activation ordering Q , which specifies the time at which nodes publicly activate. Note that for i , $q_i = u_i^* + \delta_i$, where $u_i^* = \min(t) : r_i \geq h_i$, or the first time that node i updates and has its threshold satisfied. Since i activates eventually by assumption, u_i^* exists.

Now, we construct a (G, H) for which any Q fails to be measurable. Note that the specific form of Δ does not matter, it's just important that the researcher doesn't know it. Let $V = (i, j, k, l)$ and $E = ((i, j), (i, k), (j, k), (j, l), (k, l))$. Let $h_i = 0$ while $h_j = h_k = h_l = 1$.

Node i must activate first. After it does, we have one of three cases:

1. $q_j < q_k$: When k activates, k has 2 active neighbors, so a collision has happened

2. $q_j > q_k$: When j activates, j has 2 active neighbors, so a collision has happened
3. $q_k = q_j$: When l activates, l has 2 active neighbors, so a collision has happened

We have shown that there exists a (G, H, Δ) that fails to be threshold measurable for any U when S and Δ are hidden. □

A graphical version of Theorem 2 can be found in Fig. 7.

7.2.8 Case where researcher controls node update ordering, knows activation delays, but does not know private threshold satisfaction status

We build on Theorem 2 by allowing the researcher to know node activation delays. This allows researchers to choose an update ordering that takes into account both graph structure and node activation delays in constructing the node update ordering.

Theorem 3. *Assume (G, H, Δ) are exogenous, that the researcher observes Δ but not s^t . Then the contagion process is not threshold measurable.*

Proof. Return to the graph from Theorem 2. Playing the part of nature, set $\Delta = 0$. Then the graph is not threshold measurable by logic identical to Theorem 2. Either j, k , or l will be incorrectly measured. □

7.2.9 Case where researcher controls node update ordering, knows activation delays, and knows private threshold satisfaction status

We build on Theorem 2 by allowing the researcher to observe private threshold satisfaction status s^t .

Theorem 4. *Assume (G, H, Δ) are exogenous, and that the researcher observes both Δ and s^t . Then the contagion process is not threshold measurable.*

Consider the graph in Theorem 2. Set $\Delta = 0$. Then the process is not threshold measurable by logic identical to Theorem 2.

7.2.10 Case where researcher controls node update ordering and activation delays

Theorem 5. *A contagion process is threshold measurable if (G, H) are exogenous, (U, Δ) endogenous, and the researcher observes s^t and Δ .*

Proof. For each node to be correctly measured, the contagion process must mimic an experiment in the sense that each node must update and record its status after each alter activation. Since we would like threshold measurability for arbitrary G and H , this implies that each node must update after each activation. We can construct a U and Δ that allow this.

Intuitively, we can choose node update ordering and activation delays such that at each even time period $t^{even} = \{0, 2, 4, \dots\}$ all nodes update, and at each odd time period $t^{odd} = \{1, 3, 5, \dots\}$ exactly one node publicly activates. We make use of knowing s^t to determine when individual node thresholds are satisfied apart from public activation a^t .

Let z^t be a vector indicating satisfied but unactivated nodes at t . In other words $z_i^t = 1$ if $s_i^t = 1$ and $a_i^t = 0$ for all $i \in z^t$, otherwise $z_i^t = 0$. At each even time t^{even} , have all nodes in V update simultaneously (note nothing happens for nodes who are already satisfied). For each newly satisfied node i after this update, set $z_i^t = 1$ and assign i a public activation delay $\delta_i = 2 \sum_i z_i^t - 1$. The first node to have its threshold satisfied will be assigned an activation delay of 1, the second a delay of 3, and so forth. This defines a U and Δ .

To see that such a process is threshold measurable, consider any $i \in V$. By assumption, i activates eventually. Since i updates at every time $t = \{0, 2, 4, 6, \dots\}$ and we know s^t , find $t^* = \min(t) : s_i^t = 1$. By construction, exactly one node activated publicly at $t^* - 1$, which implies that i 's true threshold is $h_i = r_i^{t^*}$, i 's exposure at t^* . \square

Importantly, we require the ability to construct U and Δ as the process is unfolding. This type of centralized control does not exist in social networks, which are decentralized by definition.

We note that there are likely other constructions of U and Δ that satisfy the measurability condition. And, as we noted above, there may be other contagion processes not discussed here that allow threshold measurability.

7.2.11 Case where researcher controls node update ordering, knows activation delays, activation delays are > 0 , and knows private threshold satisfaction status

Conjecture: this process is threshold measurable. We leave proof to future work.

7.3 Additional figures and tables

Error	Count
+0	279
+1	216
+2	104
+3	47
+4	29
+5	12
+6	10
+7	3
+8	6
+9	1
+10	3
+11	1
+14	3
+15	1
+16	1
+19	1
+20	1
+21	1
+23	1

Table 1: An inventory of the severity of overestimation for a representative simulation run. +0 error means that a node’s threshold is correctly measured, while +1 indicates that a threshold of h was measured as $h + 1$.

	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+14	+15	+16	+19	+20	+21	+23
0	12	28	11	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	29	13	4	3	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0
2	18	24	14	3	6	1	0	0	2	0	0	0	1	0	0	0	0	0	0
3	31	32	20	9	3	0	3	0	1	0	0	0	0	0	0	0	0	0	0
4	49	21	25	9	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	44	46	11	10	3	1	0	0	2	0	2	0	0	0	0	0	0	0	0
6	49	22	5	5	3	1	3	1	1	1	0	0	0	0	0	0	0	0	1
7	21	8	7	3	3	2	1	0	0	0	1	0	0	0	0	0	0	0	0
8	14	14	3	3	1	2	0	0	0	0	0	0	0	1	0	0	1	0	0
9	6	1	1	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0
10	1	2	2	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0
11	3	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
12	1	2	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
13	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Table 2: True threshold on the rows; measurement error on the columns. The first column gives the number of correctly measured nodes at that threshold, the second column gives the number of nodes with a measurement error of +1, etc. As an example, row 9 and column +7 indicates the number of nodes that have a true threshold of 9 but were measured with a threshold of $9 + 7 = 16$. Note that measurement errors are always positive.

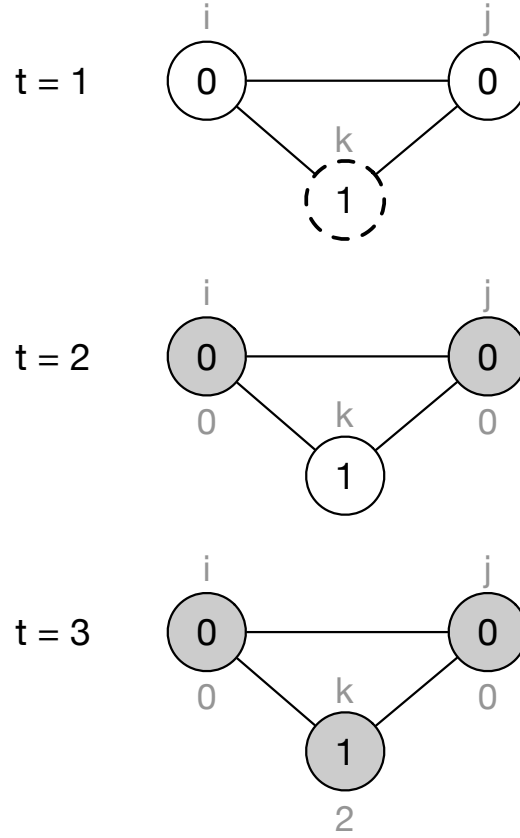


Figure 6: Graphical version of Theorem 1. Dashed lines indicate updated but non-activated nodes. Transparent nodes are inactive and did not update. Gray nodes have activated. Numbers on the node indicate the true threshold, while numbers below the node indicate the measured threshold using the exposure-at-activation-time rule. Letters above the nodes are node labels. The practical outcome of this process is that node k is measured with a threshold interval $[0, 2)$, and so is incorrectly measured.

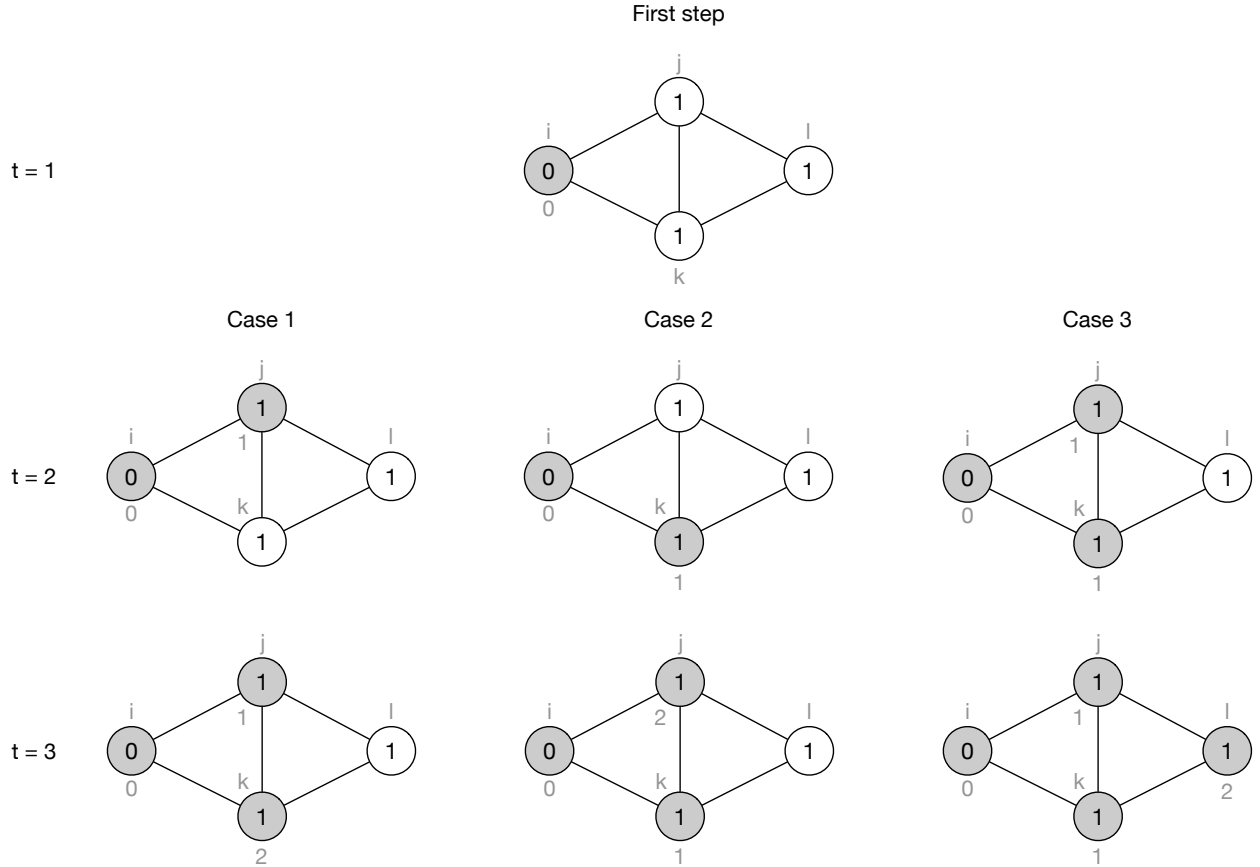


Figure 7: Graphical version of Theorem 2. Transparent nodes are inactive and did not update. Gray nodes have activated. Numbers on the node indicate the true threshold, while numbers below the node indicate the measured threshold using the exposure-at-activation-time rule. Letters above the nodes are node labels. The first step at $t = 1$ has no variation across cases and so we present it only once. We see that any update ordering produces mismeasurement of at least one node, indicated by a mismatch of the true threshold with the threshold using the exposure-at-activation-time rule.

References

- [1] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. U. S. A.*, 106(51):21544–21549, 2009.
- [2] Jordi Bascompte and Pedro Jordano. Plant-Animal Mutualistic Networks: The Architecture of Biodiversity. *Annu. Rev. Ecol. Evol. Syst.*, 38(2007):567–593, 2007.
- [3] Jonah Berger and Katherine L Milkman. What Makes Online Content Viral? *J. Mark. Res.*, 49(2):192–205, 2012.
- [4] Lawrence Blume, David Easley, Jon Kleinberg, Robert Kleinberg, and Eva Tardos. Which networks are least susceptible to cascading failures? *Proc. - Annu. IEEE Symp. Found. Comput. Sci. FOCS*, pages 393–402, 2011.
- [5] Ellsworth Campbell and Marcel Salathé. Complex social contagion makes networks more vulnerable to disease outbreaks. *Sci. Rep.*, 3:1905, 2013.
- [6] Damon Centola and Michael Macy. Complex Contagions and the Weakness of Long Ties. *Am. J. Sociol.*, 113(3):702–734, 2007.
- [7] Jie Chen, James S. Thorp, and Ian Dobson. Cascading dynamics and mitigation assessment in power system disturbances via a hidden failure model. *Int. J. Electr. Power Energy Syst.*, 27(4):318–326, 2005.
- [8] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? *Proc. 23rd Int. Conf. World Wide Web*, pages 925–936, 2014.
- [9] Andrew Chester. The Effect of Measurement Error. *Biometrika*, 78(3):451–462, 1991.
- [10] Nicholas a Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.*, 357(4):370–9, 2007.
- [11] Peter Sheridan Dodds and Duncan J. Watts. Universal behavior in a generalized model of contagion. *Phys. Rev. Lett.*, 92(21):218701–1, 2004.

- [12] P. Gai and S. Kapadia. Contagion in financial networks. *Bank Engl.*, (383):1–35, 2010.
- [13] Giacomo De Giorgi, Michele Pellizzari, Silvia Redaelli, Source American, Economic Journal, Applied Economics, No April, and De Giorgi. Identification of Social Interactions through Partially Overlapping Peer Groups. 2(2):241–275, 2010.
- [14] M. Granovetter. Threshold models of collective behavior. *Am. J. Sociol.*, 83(6):1420–1443, 1978.
- [15] Petter Holme and Beom Jun Kim. Growing Scale-Free Networks with Tunable Clustering. (2), 2001.
- [16] Xuqing Huang, Irena Vodenska, Shlomo Havlin, and H. Eugene Stanley. Cascading Failures in Bi-partite Graphs: Model for Systemic Risk Propagation. *Sci. Rep.*, 3(1219):1–8, 2013.
- [17] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. *Proc. ninth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '03*, page 137, 2003.
- [18] David Kempe, Jon Kleinberg, and Éva Tardos. Influential Nodes in a Diffusion Model for Social Networks. *Autom. Lang. Program.*, 3580:1127–1138, 2005.
- [19] A. S. Klov Dahl, J. J. Potterat, D. E. Woodhouse, J. B. Muth, S. Q. Muth, and W. W. Darrow. Social networks and infectious disease: the Colorado Springs study. *Soc. Sci. Med.*, 38(1):79–88, 1994.
- [20] Michael W. Macy. Chains of Cooperation: Threshold Effects in Collective Action. *Am. Sociol. Rev.*, 56(6):730–747, 1991.
- [21] Charles F. Manski. Identification of Endogenous Social Effects: The Reflection Problem. *Rev. Econ. Stud.*, 60(3):531, 1993.
- [22] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. *Proc. 20th Int. Conf. World wide web*, pages 695–704, 2011.

- [23] David Strang and Sarah a. Soule. Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills. *Annu. Rev. Sociol.*, 24(1998):265–290, 1998.
- [24] Steven H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.
- [25] Thomas W Valente. *Network Models of the DIffusion of Innovations*. 1995.
- [26] Thomas W Valente. Social Network thresholds in the diffusion of innovations. *Soc. Networks*, 18(95):69–89, 1996.
- [27] Thomas W Valente. Network Interventions. *Science*, 337(6090):49–53, 2012.
- [28] S van den Hof, C M a Meffre, M a E Conyn-van Spaendonck, F Woonink, He de Melker, and R S van Binnendijk. Measles outbreak in a community with very low vaccine coverage, the Netherlands. *Emerg. Infect. Dis.*, 7(3):593–597, 2001.